Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System

Alhussein, M., Ali, Z., Imran, M., & Abdul, W. (2016). Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System. *Mobile Information Systems*, *2016*, 1-12. [7805217]. https://doi.org/10.1155/2016/7805217

Link to publication record in Ulster University Research Portal

Published in:
Mobile Information Systems

Publication Status:
Published (in print/issue): 01/01/2016

DOI:
10.1155/2016/7805217

Document Version
Publisher's PDF, also known as Version of record

General rights
Copyright for the publications made accessible via Ulster University's Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact pure-support@ulster.ac.uk.

Download date: 04/05/2023

Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 7805217, 12 pages
http://dx.doi.org/10.1155/2016/7805217

Research Article

Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System

Musaed Alhussein,[1] Zulfiqar Ali,[1] Muhammad Imran,[2] and Wadood Abdul[1]

[1]*Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia*
[2]*Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia*

Correspondence should be addressed to Zulfiqar Ali; zuali@ksu.edu.sa

Received 31 December 2015; Accepted 3 April 2016

Academic Editor: Mehmet Orgun

Copyright © 2016 Musaed Alhussein et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

An automatic gender detection may be useful in some cases of a mobile healthcare system. For example, there are some pathologies, such as vocal fold cyst, which mainly occur in female patients. If there is an automatic method for gender detection embedded into the system, it is easy for a healthcare professional to assess and prescribe appropriate medication to the patient. In human voice production system, contribution of the vocal folds is very vital. The length of the vocal folds is gender dependent; a male speaker has longer vocal folds than a female speaker. Due to longer vocal folds, the voice of a male becomes heavy and, therefore, contains more voice intensity. Based on this idea, a new type of time domain acoustic feature for automatic gender detection system is proposed in this paper. The proposed feature measures the voice intensity by calculating the area under the modified voice contour to make the differentiation between males and females. Two different databases are used to show that the proposed feature is independent of text, spoken language, dialect region, recording system, and environment. The obtained results for clean and noisy speech are 98.27% and 96.55%, respectively.

1. Introduction

The applications of automatic gender detection (AGD) system have increased significantly due to the recent developments in speech/speaker recognition, human-computer interaction, and biometric security systems including authentication to access data, surveillance, and security. Gender detection systems limit the search of an imposter to half of the space in many recognition and security systems, where the ultimate goal is the identification of a person. Considering different feature extraction and modeling techniques, an AGD for recognition and security systems should be implemented in such a way that it should not increase the complexity of the whole system. Moreover, a gender detection system can be used for automatic transfer of a phone call of a male/female to the relevant person or department. Furthermore, the accuracy of gender dependent models is higher than gender independent models [1].

In a mobile healthcare system [2–5], automatic gender detection can play a significant role. There are some vocal folds pathologies [6, 7], which are biased to a particular gender; for example, vocal folds cyst can be seen particularly in female patients [8, 9]. If there is a mechanism to automatically detect the gender of the patient, it is easier for a care giver or a healthcare professional to prescribe the appropriate treatment. In this system, the voice or speech of the patient is recorded via a smart device, which is connected to the Internet. The voice or speech is then transmitted to a cloud, where a cloud manager authenticates the patient. The manager distributes the task of feature extraction and classification to various servers, where a decision of gender is made. The decision along with medical data is transmitted to registered healthcare professionals for proper treatment.

In most of the studies [10–16], the acoustic features used for the gender detection depend on the accurate estimation of the fundamental frequency. The accurate estimation of the

fundamental frequency is itself a challenging task. Inaccurate estimation of the fundamental frequency may lead to a significant reduction in the accuracy of a gender detection system. Moreover, various traditional speech features such as linear predictive coefficients (LPC), linear predictive cepstral coefficients (LPCC), Mel-frequency cepstral coefficients (MFCC), perceptual linear predictive coefficients (PLP), and relative spectral PLP coefficients (RASTA-PLP) are used in [10, 12–14, 17, 18] for gender detection. The author in the study [19] claimed that the features used for speech recognition may not provide good results for gender detection. Therefore, it is necessary to explore new features for the gender detection other than traditional speech features, and those features should not rely on the accurate estimation of the fundamental frequency.

In this paper, we proposed a new type of feature for automatic gender detection. The proposed feature considers a speech signal of male and female speakers in time domain and provides a single value in the form of area under the modified voice contour (MVC). The proposed feature does not depend on the estimation of the fundamental frequency, and it has provided good results as compared to the existing features.

Automatic gender detection system based on different types of feature and classifier with varying accuracies are reported in the literature. The acoustic characteristics of humans are based on gender due to physiological changes in glottis, vocal tract thickness, and length. Therefore, researchers are trying to find out the most discriminative features for gender detection. For example, two acoustic features, pitch and first formant, are extracted by linear predictive analysis to construct a gender detection system in [17]. The first feature relates to voice source and the second to the vocal tract. The pitch and the formant frequencies of females are higher as compared to those of males. Euclidean distance and nearest neighbor based classifier has been implemented to detect genders. In the study of Wu and Childers [10], a number of the acoustic parameters, such as autocorrelation, linear prediction, cepstrum, and reflection, extracted from vowels, and voiced and unvoiced fricatives, performed well for gender detection. The study concludes that, for a given gender, the information is time invariant, phoneme independent, and speaker independent. In [11], an accuracy of 96% is attained when pitch is inputted to Multi-Layer Perceptron (MLP) neural network. Pitch, energy, and 12-dimensional MFCC are fed to Support Vector Machine (SVM), and the performance with the gender detection system is 95% [12]. To perform gender detection using vocal source, different vocal source parameters are extracted, and detection rate of 94.7% for male and 95.5% for female voice is achieved in [20].

A comparison between various cepstral features is provided in [13] when extracted from voiced frames only and from a running speech. The cepstral features, MFCC, LPCC, and PLP, are used with their delta coefficients to perform the experiments under three different conditions. Sigmund [18] used selected MFCC to classify the male and female by using short segments of vowels as well as sentences.

A robust gender detection system is developed by Zeng et al. in [14]. The developed system has been tested for the noisy environment, and dependency of the language is also considered for the evaluation of the system. The obtained accuracy is 95%, which shows the robustness of the developed system against noise. The experiment shows that the system is independent of language as well. Relative spectral PLP features and pitch of the male and female speakers are used for gender detection. Chen et al. [15] proposed a gender detection system for children of two age groups of 8-9 years and 16-17 years. The obtained accuracies are 60% and 94%, respectively, for two age groups. Different acoustic features, source spectral magnitude, cepstral peak prominence, and harmonic-to-noise ratio, are used to implement the system. Sedaaghi [21] conducted a comparative study for the gender detection by using two different databases. Various classifiers and acoustic features are used in [16] for gender classification system, and the best reported accuracy is 95%. A total of 113 features are used in the study and Bayes classifier is used for feature selection. The features are grouped into pitch, formant, spectral, and intensity. K-nearest neighbor, SVM, artificial neural networks, and Gaussian mixture model (GMM) are used as classifier in [16]. An automatic gender detection system for Hindi speech is developed in [19] by using MFCC as features and Euclidean distance as a classification method. The authors have mentioned in this study that the same features can be used for gender and speech recognition. However, use of same features for both recognition systems cannot guarantee good performance.

An initial investigation of the proposed feature was done in [22] when MVC was used to measure the voice intensity of the speech sample to discriminate between the genders. The database used was Arabic digits and manual threshold was used to classify the males and females. To authenticate the performance of the proposed feature, TIMIT database was considered and automatic classification of the genders was done by the SVM in [23].

In this study, we have investigated the proposed feature in many aspects, which makes this study different than [22, 23]. The most important factor is to observe the robustness of the proposed feature against noise. Background/environmental noise in speech based applications can degrade the performance of the developed system and, therefore, cannot be neglected. Hence, a white noise at different sound-to-noise ratio (SNR) levels is added in the speech signal of both genders and, then, performance of the developed system is evaluated. The results with clean speech are also obtained to make a comparison between the results of clean and noisy speech. Moreover, many experiments are performed to show that the proposed feature is independent of the language, text, and recording systems. Two databases are used for this purpose; the first database is in English, and the second is in Arabic. Both databases are recorded by using different recording systems, and spoken text is also different. Furthermore, the results of the proposed feature are compared with the pitch plus RASTA-PLP features which provided the best accuracy of 95% in [14]. The proposed feature provided good accuracy and can be used with speaker recognition and biometric security systems to reduce the system's complexity by splitting the search space into two halves.

The rest of the paper is organized as follows: Section 2 describes our proposed automatic gender detection system.

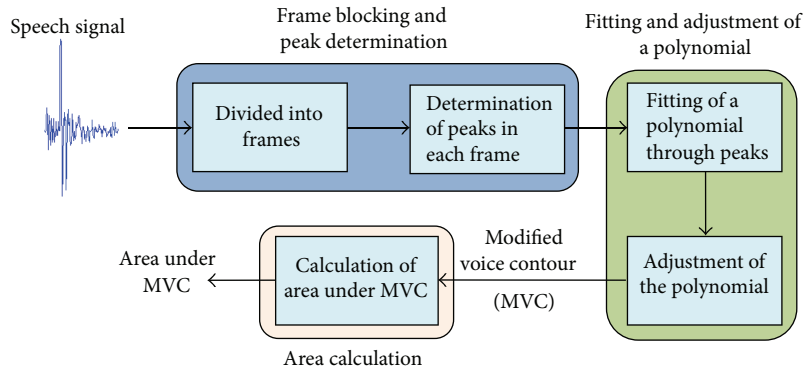


Figure 1: Block diagram to determine modified voice contour and area under it.

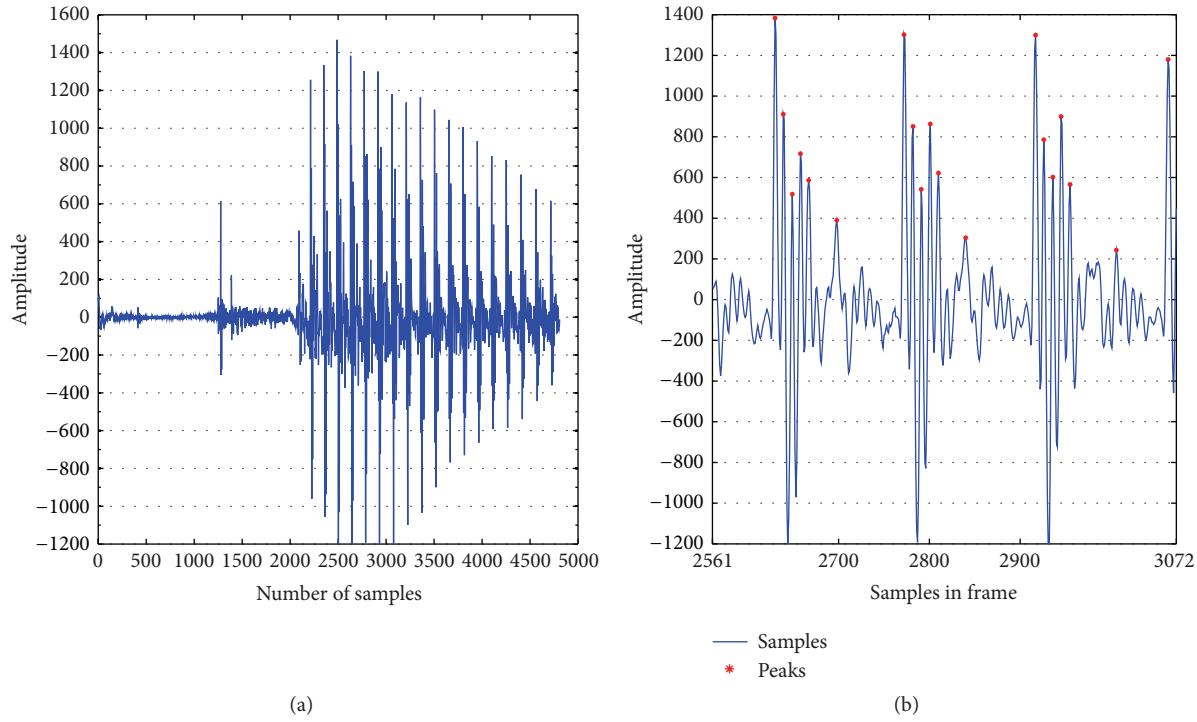


Figure 2: (a) A speech signal. (b) Peaks in a frame.

Section 3 provides the description of speech databases. Section 4 explains the experimental setup and results of the proposed and existing AGD systems. Section 5 analyzes the results and conclusions are drawn in Section 6.

2. Proposed Automatic Gender Detection System

In this study, an automatic gender detection system by using the proposed feature is developed. The proposed feature determines the voice intensity of a speech signal by using the MVC. To implement the feature, Simpson's rule is used to calculate the area under the MVC. The MVC is obtained after adding a factor in a polynomial of degree three that is fitted through the peaks. The peaks are found from each frame when a speech utterance is blocked into frames. At the end, the calculated area is fed to SVM to make the decision about the type of gender. A block diagram to determine the MVC of a speech signal is shown in Figure 1. The implementation of the proposed feature is divided into five major components and they are grouped in three steps: (1) frame blocking and peak calculation, (2) fitting and adjustment of a polynomial, and (3) calculation of area under the MVC by using Simpson's rule. To make a decision about the gender, a binary classifier SVM is used.

2.1. Frame Blocking and Peak Determination. The speech signal, as shown in Figure 2(a), is recorded at the sampling frequency of 16 KHz. A speech signal can be considered dynamic in nature because it changes with time. Therefore,

analysis for whole speech is not possible due to variation in a speech signal. It is the reason that speech may be divided into small frames ranging within 10~40 milliseconds. The variation in size of a frame from 10 to 40 milliseconds is not critical; however, a frame of size 32 milliseconds has provided slightly better results than a frame of lengths 16 and 25 milliseconds [24].

To determine the MVC, peaks are found after blocking the whole speech signal into frames. The length of each frame is 32 milliseconds, and it contains 512 samples. The peaks higher than a certain threshold value thresh are determined in each frame. The thresh is calculated for the whole speech signal by using (1) and kept the same for all the frames of that signal. A frame showing the calculated peaks is depicted in Figure 2(b):

$$\text{thresh} = \left(\text{amp}_{\text{Max}} - \text{amp}_{\text{Min}}\right) * 0.1 + \text{amp}_{\text{Min}}, \quad (1)$$

where amp_{Min} and amp_{Max} are 3% and 97% percentiles of the amplitude in a speech signal, respectively. The thresh varies from signal to signal. Different words exhibit different patterns of the waveform, and, hence, amplitude in a speech signal is also varied. Therefore, to calculate the thresh automatically for each speech signal, (1) is implemented. The relation provided in (1) has also been used successfully in other applications to determine the threshold [25].

2.2. Fitting and Adjustment of the Polynomial. After the calculation of the peaks in each frame, they are joined together. Then, a polynomial of degree three, $g(x)$, is fitted through these peaks to form a curve as shown in Figure 3; the curve is composed of diamonds. It can be observed from Figure 3 that the fitted polynomial passes through the peaks points, hence, not making an envelope over the joined peak points. Therefore, a factor given by (2) is added in the polynomial $g(x)$ to get the MVC:

$$\text{factor} = \left(\max\left(\text{peaks}\right) - \max\left(g\right)\right) * 0.70, \quad (2)$$

where "peaks" is a vector containing peaks of all frames and g is a vector containing all points on the fitted polynomial $g(x)$. Equation (2) provided a factor to adjust the fitted polynomial $g(x)$ which is equivalent to the 70% of the difference between the highest peak and the maximum point on $g(x)$. To avoid the biased peaks, the 70% of the difference is considered; otherwise, the adjusted curve will be misfitted and will not make an envelope over the peaks. After adding the factor in the fitted polynomial, the MVC composed of "$*$" is shown in Figure 3 and is obtained as

$$\text{MVC} = g\left(x\right) + \text{factor}. \quad (3)$$

2.3. Area Calculation. Finally, to obtain the voice intensity, the area under the MVC is calculated with Simpson's rule of numerical integration [26–28], given by (4). The rule divides the region under the MVC into trapeziums, as shown in Figure 4, and calculates area for each trapezium. Then, it takes

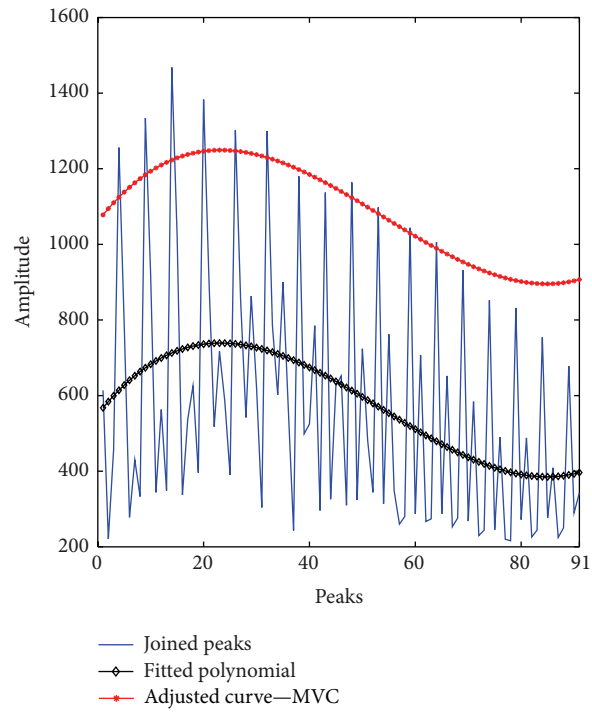


Figure 3: The modified voice contour (MVC).

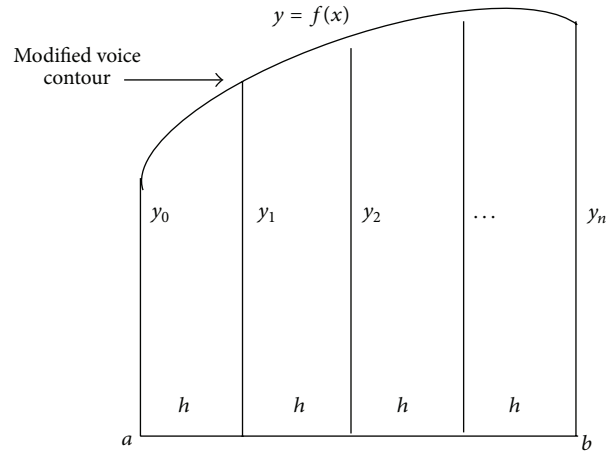


Figure 4: Area under the curve by Simpson's rule.

summation of the area of all trapeziums to provide the total area under the MVC. Simpson's rule calculates area as follows:

$$\text{Area} = \int_a^b f\left(x\right) dx$$

$$\approx \frac{h}{3}\left(y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + y_n\right), \quad (4)$$

$$h = \left(\frac{b-a}{n}\right),$$

where a and b are the first and the last points on the MVC.

The number of trapeziums under MVC is represented by n. In Simpson's rule, good approximation for area under

a curve can be achieved by increasing n because error in approximation of area decreases as the number of trapeziums increases. The consideration of large value for n is also not feasible because it will increase the computational cost. The value of n is always even in Simpson's rule and it is set to 50 in this study after trying $n = 20, 50, 100$, and 150.

2.4. Support Vector Machine. SVM was proposed by Vapnik [29] and it becomes popular due to its good performance and low computational cost as compared to other classification techniques such as GMM and Hidden Markov Model (HMM). In the developed AGD system, the SVM takes the area under the MVC to make the decision about gender of the speaker. SVM constructs a decision surface (hyper plane) to maximize the distance between two classes, a positive class and a negative class [30]. The dimension of hyper plane depends on the dimension of feature vector given to SVM.

In this study, SVM is implemented by using LIBSVM [24] with a radial basis function (RBF) as kernel, given by

$$K\left(x, x'\right) = \exp\left(-\gamma \left\| x - x' \right\|^2\right), \tag{5}$$

where x is the training sample, x' is the testing sample, and γ is a free parameter. SVM is a linear classifier; however, in most of the cases, data is not linearly separable. Therefore, kernel function is implemented to map the original input space to higher dimensional space, where features are lineally separable. During implementation, male speakers are represented as a positive class and female speakers are represented as a negative class.

3. Material

To evaluate the proposed feature independent of the text and language, two different databases are used. The language of the first database is English, and that of the second is Arabic. Both databases are recorded by using different recording systems and environments, and the spoken text is also different in them.

3.1. TIMIT Database. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) [31] is used to perform the experiments with English language. The database contains 630 speakers of eight different dialect regions of the United States. Each speaker has recorded 10 sentences at sampling rate of 16 KHz by a condense microphone, where sentence 1 and sentence 2 are the same for all speakers. These fixed sentences are as follows.

Sentence 1. She had your dark suit in greasy wash water all year.

Sentence 2. Do not ask me to carry an oily rag like that.

Only one utterance of these sentences is available because each speaker has recorded these sentences only once. Database includes speakers of many dialect regions but, in this study, we included only those dialect regions in which the total number of speakers is about 100 and contained at least 30 female speakers. Numbers of speakers in dialect regions 2, 4, and 5, labeled as D2, D4, and D5, are 102, 100, and 98, respectively, while the numbers of male and female speakers in each dialect region are listed in Table 1.

Table 1: Number of male and female speakers in different dialect regions.

Dialect regions	Label	Number of speakers		
		Male	Female	Total
2	D2	71	31	102
4	D4	69	31	100
5	D5	62	36	98

Table 2: List of selected words.

Sentence	Word position	Word
1	4	Dark
	5	Suit
	7	Greasy
	8	Wash
	9	Water
	10	All
	11	Year
2	2	Ask
	5	Carry
	7	Oily
	8	Rag
	9	Like

Twelve words, randomly selected, are extracted from sentences 1 and 2. So the total number of available samples for experimentation is 3600 (= 300 × 12). The list of the extracted words is presented in Table 2. The second column provides the position of the word in the sentence. For instance, the first entry of the second column represents that "Dark" is at fourth position in sentence 1.

3.2. The Arabic Database. The Arabic database [32] contains speech samples of 71 speakers: 53 males and 18 females. Each speaker has recorded one utterance of each Arabic digit from one to nine, as listed in Table 3. Speakers recorded the digits with a professional side-address condenser microphone (SHURE PG42) connected to a high-quality mixer (Yamaha MW12CX) at sampling rate of 16 KHz.

4. Experimental Setup and Results

Various experiments are performed to investigate the performance of the proposed feature by using English and Arabic database. Numbers of male and female speakers are not the same in the selected regions D2, D4, and D5 in the English database. So, firstly, one experiment for each dialect region is performed to check the biasness of the proposed feature for gender imbalanced corpus. Secondly, few experiments are performed after making the corpus balanced. Thirdly, noise of different SNRs is added to investigate the robustness of the

Table 3: List of Arabic digits and their IPA.

Digits			
Symbol	In roman English	In Arabic	IPA
1	Wahed	واحد	wa:- ħid
2	Athnayn	أثنين	ʔiθ-ni:n
3	Thalathah	ثلاثه	θa-la:- θah
4	Arbaah	أربعه	ʔar-ba-'ah
5	Khamsah	خمسه	xam-sah
6	Setah	سته	Sit-tah
7	Sabaah	سبعه	Sab-'ah
8	Thamanyah	ثمانية	θa-ma-ni-jah
9	Tesaah	تسعه	tis-'ah

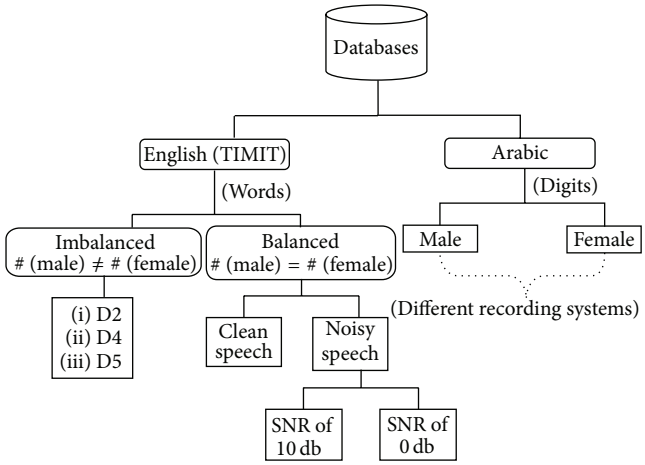


Figure 5: The experimental setup.

proposed feature against noise. Finally, some experiments are performed with Arabic database to observe the accuracy for another spoken language. The text recorded by the speakers in English and Arabic databases is different. The experimental setup is depicted in Figure 5.

All results for gender detection are obtained by using the 5-fold approach. In the 5-fold approach, the database is divided into five distinct subsets. Each time, one of the subsets is used for testing, and the remaining four subsets are used for training of the system. The performance of the proposed AGD system is carried out by using the following parameters:

True positive (TP): the male speaker detected by the system as a male.

True negative (TN): the female speaker detected by the system as a female.

False positive (FP): the female speaker detected by the system as a male.

False negative (FN): the female speaker detected by the system as a male.

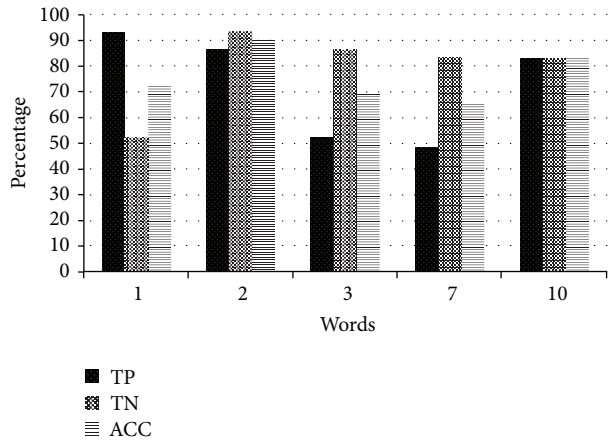


Figure 6: Comparison of performance measures for different words.

Sensitivity (SE): the likelihood that the system detects male when the input is male speaker,

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{6}$$

Specificity (SP): the likelihood that the system detects female when the input is female speaker,

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{7}$$

Accuracy (ACC): the ratio between correctly detected files of the genders and the total number of files,

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100. \tag{8}$$

Area under curve (AUC): the area under the Receiver Operating Characteristic (ROC) curve.

4.1. Gender Detection with Imbalanced Corpus. In this section, the experiments are performed to observe the accuracy of the proposed AGD system by using each word individually and all words simultaneously. In all experiments, numbers of male and female speakers are different.

4.1.1. Gender Detection by Using Individual Words with Imbalanced Corpus. The results of gender detection for each word listed in Table 2 are summarized in Table 4. The dialect region D5 is used in these experiments as it has more number of females as compared to D2 and D4. The highest result for TP is 93% for word 1, and for TN it is also 93% but for word 2. The maximum achieved accuracy is for word 2 and it is 90%. Table 4 provides the analysis of all words in terms of different performance metrics so that we may observe the contribution of each word to gender detection. A comparison of different words having good results in terms of TP, TN, and ACC is depicted in Figure 6.

Table 4: Results for gender detection with individual words by using the proposed feature.

Performance measures	Words											
	1	2	3	4	5	6	7	8	9	10	11	12
TP (%)	93	86	52	72	72	72	48	55	72	83	59	41
TN (%)	52	93	86	48	41	48	83	69	69	83	69	66
FP (%)	48	7	14	52	59	52	17	31	31	17	31	34
FN (%)	7	14	48	28	28	28	52	45	28	17	41	59
ACC (%)	72	90	69	60	57	60	66	62	71	83	64	53

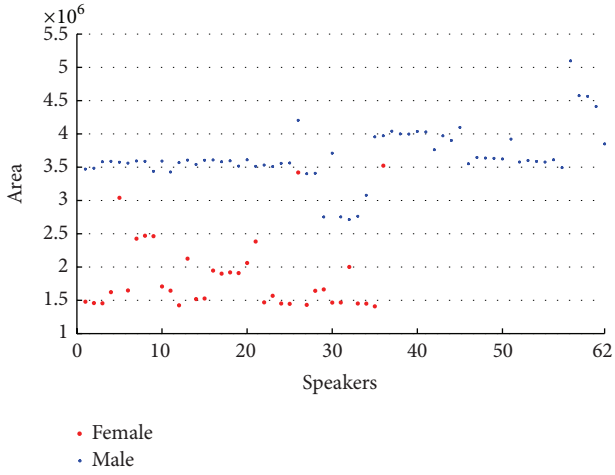


Figure 7: Plot of area under MVC for Word 9 of D5.

4.1.2. Gender Detection by Using All Words Simultaneously with Imbalanced Corpus. Area of all words for each speaker is calculated and, then, fused before providing to SVM for gender detection. Twelve-dimensional feature vector is used for each speaker where each dimension represents an area of the word. The accuracy of 96% is achieved, where obtained SE and SP are 94% and 100%, respectively. Twenty-seven times, out of 28, system detects the gender correctly: 17 out of 18 for males and 10 out of 10 for females. The plot of area under the MVC for male and female speakers for word 9 of dialect region D5 is depicted in Figure 7. Dialect region D5 has 98 speakers: 36 females and 62 males.

By analyzing the results of gender detection by using words individually and all words simultaneously, it can be inferred that it is better to use more than one word to achieve high detection rate. Therefore, in the rest of the experiments, we will use all words simultaneously to achieve better gender detection rate. With all words listed in Table 2, the accuracy of 96% is achieved for D5. The number of male and female speakers in that experiment was 62 and 36, respectively.

It might be assumed that the proposed feature is biased when numbers of samples are different for male and female speakers. To provide the answer about biasness of the proposed feature, it is necessary to use the equal number of male and female speakers for the training and testing of the system. The only available option is to include the female speakers of other dialect regions. It will make a balance between both genders and it will increase the total number of speakers.

Table 5: Results of gender detection by using all words simultaneously for all dialect regions.

Dialect region	Number of speakers (M + F)	Accuracy
D2	71 + 31	94%
D4	69 + 31	96%
D5	62 + 36	96%

M and F stand for male and female speakers, respectively.

However, before doing so, two more experiments for the dialect regions D2 and D4 are performed to investigate the effect of dialect regions. The obtained detection rates for D2 and D4 are 94% and 96%, respectively, which show that dialect regions do not affect the accuracy of the developed system and the performance of the proposed feature is good. Hence, speaker of different dialect regions can be grouped to make the number of male and female speakers equal. The accuracy for D2, D4, and D5 is mentioned in Table 5.

4.2. Gender Detection with Balanced Corpus. It is concluded in Section 4.1 that use of all words simultaneously provided good gender detection rate. Hence, we will continue with it in the rest of the experiments. In addition, the results in Table 5 show that the proposed feature is independent of dialects. Therefore, to make a balance between the numbers of males and females, we can combine speakers of different dialects, and the balanced corpus will be used in the rest of the experiments in this section.

Numbers of female speakers in the dialect regions D2, D4, and D5 are 31, 31, and 36, respectively, and the same numbers of males are taken from these regions to make a balance between male and female speakers. Now, the total number of females is 98 (= 31 + 31 + 36) and the same number of male speakers makes the total number of speakers equal to 196.

A white noise of SNR of 10 db and 0 db is added to the balanced corpus to check the robustness of the proposed feature against noise, and the obtained results are compared with the existing system presented in [14]. The pitch and RASTA-PLP [33] were extracted in [14] from the clean and noisy speech, and eight Gaussians were considered to construct GMM. In this study, to determine the optimized parameters of GMM, mean, covariance matrix, and prior probability, the Expectation-Maximization (EM) algorithm [34] is implemented, while these parameters are initialized by using k-means algorithm [35]. In the GMM based gender detection system, a GMM for both genders is developed. A

Table 6: Results of gender detection for existing and proposed systems with clean speech.

Performance measures	Existing system			Proposed system
	4 Gaussians	8 Gaussians	16 Gaussians	
TP (%)	99.13	99.71	99.71	100.00
TN (%)	91.66	91.66	95.11	96.55
FP (%)	8.33	8.33	4.89	3.45
FN (%)	0.86	0.29	0.29	0.00
SE	0.99	1.00	1.00	1.00
SP	0.92	0.92	0.95	0.97
ACC (%)	95.4	95.68	97.41	98.27
AUC	0.9510	0.9734	0.9795	0.9845

test utterance, for detection of gender of a speaker, will be compared with both models. The model that has a maximum likelihood with the test utterance will be the gender of that test utterance.

4.2.1. Gender Detection for Clean Speech with Balanced Corpus. Two experiments are performed to observe the behavior of the proposed feature for clean speech. In the first experiment, 13 coefficients are extracted per frame in each word. The first coefficient is pitch, and the rest of the twelve features are 11th-order RASTA_PLP coefficients. The extracted features are inputted to the GMM to construct the genders' models by using 4, 8, and 16 Gaussians for male and female detection. The first experiment represents the existing AGD system presented in [14]. This experiment is performed to compare the results with our proposed features.

In the second experiment, the proposed feature provides one value (area under the MVC) for one word which makes the proposed system more efficient. Then, calculated area under the MVC is fed to SVM to make the decision about the gender type. The results of both experiments are provided in Table 6.

The accuracy of the proposed feature is 98.27%, which is better than the existing system. The proposed feature dominates in all performance parameters. The true male detection rate is 100%, and for female, it is 96.55%. Only 3.45% of the female speakers are detected as male, while no male speaker is recognized as female. The experiment for the existing system is performed with the different number of Gaussians to find the best detection rate for that system.

The ROC curve for each system is plotted to analyze their performance, as shown in Figure 8. False positive rate (1 – specificity) and true positive rate (sensitivity) are taken along x-axis and y-axis, respectively. All unique numbers in the decision values of SVM are considered as cut-off points to draw the curve accurately. For existing system, the decision values of the highest accuracy, that is, 97.41%, are used to plot the curve. The area under the ROC curve for the proposed feature is greater than the existing system.

4.2.2. Gender Detection for Noisy Speech with Balanced Corpus. To observe the behavior of the proposed feature in noisy environment, a number of experiments are performed with the speech containing white noise of SNR of 10 db and 0 db, and results are compared with the existing system. The results of the systems for SNR of 10 db and 0 db with different performance parameters are summarized in Tables 7 and 8, respectively.

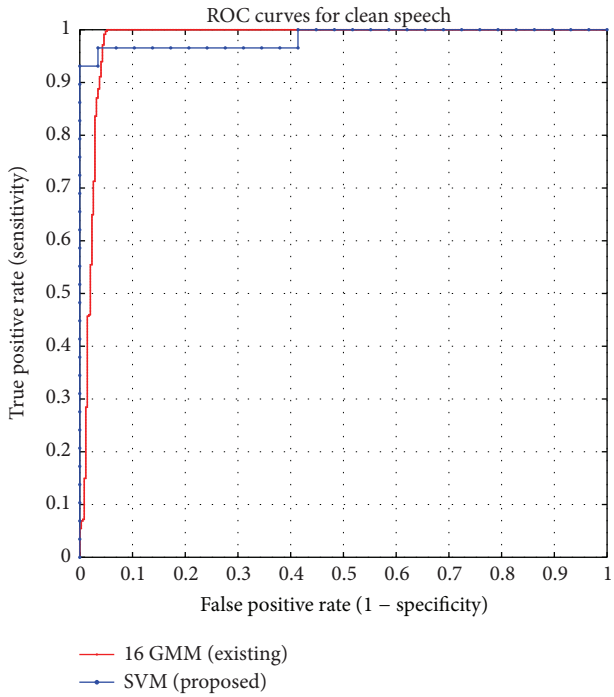


Figure 8: ROC curves of existing and proposed systems for clean speech.

By observing the obtained results, it can be inferred that the feature obtained by calculating area under the MVC can perform well even in noisy environments for the gender detection, and, again, it dominates the RASTA-PLP in most of the performance parameters. The proposed feature has provided accuracies of 96.55% and 94.82% for the SNR of 10 db and 0 db, respectively, which are better than the existing system. The ROC curves for both SNRs for existing and proposed feature are depicted in Figures 9 and 10.

4.3. Gender Detection with Arabic Corpus. To endorse the truths about the proposed feature that it can achieve good

Table 7: Results of gender detection for existing and proposed systems with SNR of 10 db.

Performance measures	Existing system			Proposed system
	4 Gaussians	8 Gaussians	16 Gaussians	
TP (%)	89.37	97.13	96.84	100.00
TN (%)	83.33	89.66	91.09	89.65
FP (%)	16.67	10.34	8.91	10.34
FN (%)	10.63	2.87	3.16	0.00
SE	0.89	0.97	0.97	1.00
SP	0.83	0.90	0.91	0.90
ACC (%)	95.25	93.96	96.12	96.55
AUC	0.9612	0.9732	0.9721	1

Table 8: Results of gender detection for existing and proposed systems with SNR of 0 db.

Performance measures	Existing system			Proposed system
	4 Gaussians	8 Gaussians	16 Gaussians	
TP (%)	89.37	97.13	96.84	100.00
TN (%)	83.33	89.66	91.09	89.65
FP (%)	16.67	10.34	8.91	10.34
FN (%)	10.63	2.87	3.16	0.00
SE	0.89	0.97	0.97	1.00
SP	0.83	0.90	0.91	0.90
ACC (%)	86.35	93.39	93.96	94.82
AUC	0.8867	0.9645	0.9609	0.9893

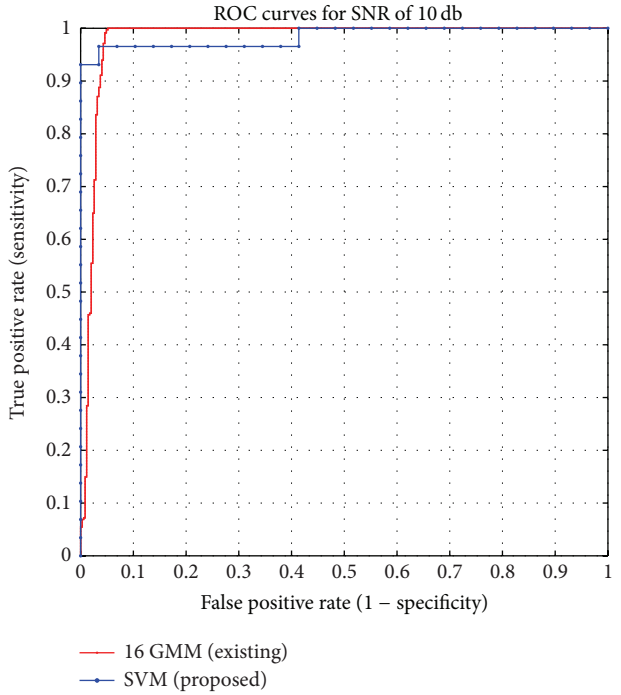


Figure 9: ROC curves of existing and proposed systems for noisy speech 10 db.

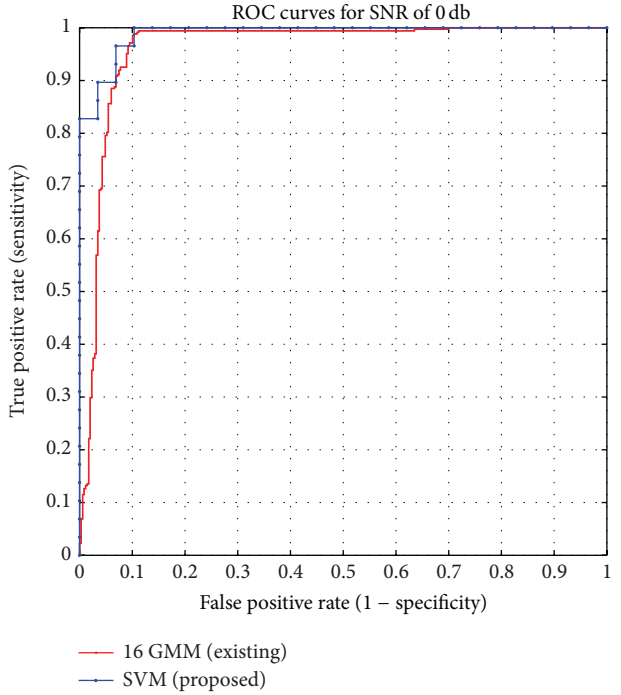


Figure 10: ROC curves of existing and proposed systems for noisy speech of 0 db.

detection rate for other spoken languages and it is independent of text and recording equipment, few experiments are performed. This investigation is performed by using Arabic digits from one to nine, listed in Table 3, uttered by 53 male and 18 female speakers. The area of the MVC is measured by

Table 9: Results of gender detection for existing and proposed systems with Arabic digits.

Performance measures	Existing system			Proposed system
	4 Gaussians	8 Gaussians	16 Gaussians	
TP (%)	92.5	94.3	96.2	100
TN (%)	88.9	94.4	94.4	100
FP (%)	11.1	5.6	5.6	0
FN (%)	7.5	5.7	3.8	0
SE	0.92	0.94	0.96	1.0
SP	0.88	0.94	0.95	1.0
ACC (%)	91.5	94.3	95.7	100
AUC	0.93	0.95	0.97	1.0

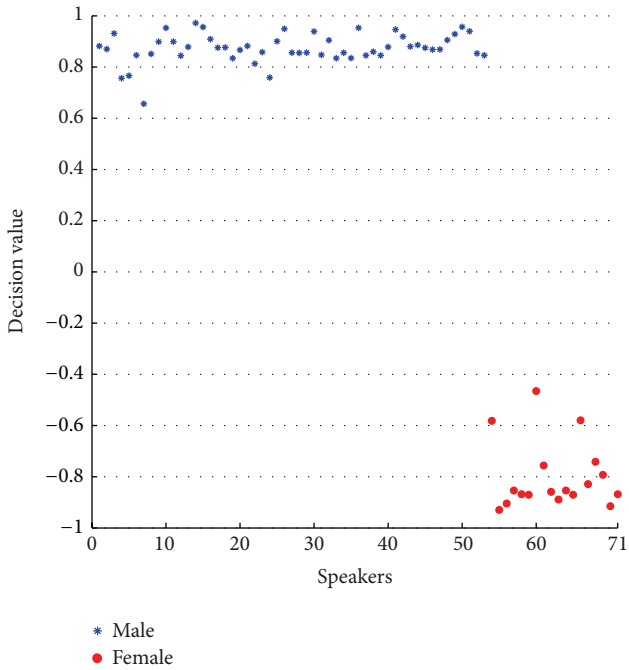


Figure 11: Decision values of the genders for Arabic digits obtained by SVM.

following the steps depicted in Figure 1 and given to SVM for detection of the gender.

Area of each digit is calculated for every speaker, so the dimension of feature vector for each speaker is nine when it is fed to SVM. The decision values obtained, during the classification of genders, by SVM are plotted in Figure 11. The area under the MVC of one or two utterances could be plotted easily, but in this experiment, areas of nine digits are fused. The interpretation of such multidimensional features is not easy to understand by a human mind. Therefore, studies based on multidimensional analysis introduce machine learning stage in the systems [36], and decision values obtained from the classifier can be considered as discriminate measures between classes. Hence, in Figure 11, we have plotted the decision values obtained from SVM.

It can be observed from Figure 11 that the values for the positive and negative classes are perfectively classified. There is no room for the confusion of the genders with each other as they are separated by good margin. The obtained accuracies for the genders are 100%. TP and TN are also 100% while FN and FP are 0% for both male and female. The results of the existing system and proposed system are provided in Table 9. The performance of the proposed system is much better than that of the existing system.

5. Discussion

A noise robust AGD system by using the proposed feature is developed in the study. The proposed feature depends on the voice intensity of a speech signal. In human voice production system, air pressure generated by lungs passes through the windpipe, also known as a trachea. The generated pressure vibrates the vocal folds which reside right on the top of the trachea. The vibration of the vocal folds, open and close, produces a voice which travels through human's mouth and generates speech. The characteristics of the voice vary from person to person due to varying shape, length, and thickness of the vocal folds. Therefore, voices of people feel significantly different from each other. The length of the vocal folds for a human usually lies between 12 and 24 millimeters (mm), whereas the thickness is 3 to 5 mm [37].

The size of the vocal folds also depends on the gender of human beings. The vocal folds length for female is approximately from 12.5 mm to 17.5 mm, while for a male, the length is from 17.5 mm to 24 mm [38]. Due to longer vocal folds, the pitch of male's voice becomes lower, and, therefore, the voice of a male feels heavier than that of a female. The heavy voice contains more voice intensity, and it is the main motivation to propose a new type of feature for gender detection. The proposed feature measures the voice intensity of the speech signal by calculating area under the MVC. It can be seen in Figure 7 that calculated area under MVC for male speakers is larger than that for a female speaker because the voice of a male has more intensity than a female speaker.

The proposed feature does not rely on the accurate estimation of the fundamental frequency which is itself a difficult task. Most of the acoustic features such as formants, harmonic-to-noise ratio, and pitch estimation depend on accurate estimation of the fundamental frequency. If fundamental accuracy is not determined accurately, the systems based on such a type of features may affect the results. In study

[19], the author claimed that traditional speech features may not perform well in a gender detection system. Therefore, a new type of feature is proposed in this study for automatic gender detection.

The developed AGD system has been evaluated in different ways by using clean speech, noisy speech, balanced speech corpus, imbalanced speech corpus, two spoken languages, and different recording systems/environments and text.

For the gender imbalanced corpus, the obtained detection rates are in the range of 94% to 96% for the D2, D4, and D5. After making the gender corpus balance by including the female speakers of the three dialects, the true positives (TP) are 100% and true negatives (TN) are 96.55% for the proposed feature. For the existing system, the best TP and TN are obtained with 16 Gaussians, and they are 99.71% and 96.55%, respectively. The accuracy and the area under the ROC curve for the proposed feature are also better than those for the existing system.

The gender detection rate of the proposed system for noisy speech is also higher than that of the existing system. The TP and TN for both SNRs of 10 db and 0 db are 100% and 89.65%, respectively. For existing system, the obtained TP is 97.13% with 8 Gaussians, and the TN is 91.09% with 16 Gaussians for both SNRs. The area under the ROC curve for the proposed feature is 2.84% better than that for the existing system for 10 db and 3.45% for 0 db.

The proposed feature has provided promising result with the Arabic digits. All males and females are perfectly classified by the feature and obtained detection rate is 100%. The results also show that the feature is capable of detection of genders with any spoken language. Overall, the proposed feature performed well under different circumstances. The word by word experiments for gender detection show that proposed feature can help in finding the words and phonemes that can provide good detection rate.

To observe the statistical significance, Mann-Whitney U test is performed at 5% significant level. The obtained p value for clean speech is $0.10E-5$, for 0 db noise is $0.12E-4$, and for 10 db noise is $0.15E-4$. All p values are less than 0.05 which reject the null hypothesis that decision values of male and female speakers are from continuous distribution with equal medians. The Mann-Whitney U test shows that the proposed system can differentiate between males and females significantly.

6. Conclusion

A new type of feature for the gender detection system is proposed in this study. The developed system can be used in the mobile healthcare systems as it provided good detection rate in both normal and noisy environments. The use of the proposed system with the mobile healthcare systems may assist the doctors in assessing and prescribing appropriate medication to the patient.

The proposed system determines the MVC of the speech signal and then finds area under the MVC to differentiate between male and female speakers. The area under the MVC represents the voice intensity of a speaker. The voice intensity of a speaker is highly dependent on the vocal folds. The size of the vocal folds in a male speaker is longer than that in a female speaker which makes the voice of a male heavy. Therefore, male speakers have more intensity in their voices than females.

Many experiments are performed by using two databases of different languages to evaluate the proposed method and to test its validity under different circumstances. With the help of conducted experiments, we can conclude that the proposed feature can perform equally well in any language. It is unbiased and independent of language, spoken text, and recording equipment. Moreover, the proposed feature is able to provide good detection rate even for the noisy environment. All obtained results are better than the existing AGD system.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University, Riyadh, Saudi Arabia, for funding this work through the research group Project no. RG-1436-016.

References

[1] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications," *Journal of Intelligent Information Systems*, vol. 24, no. 2, pp. 179–198, 2005.

[2] M. Shamim Hossain and G. Muhammad, "Cloud-assisted Industrial Internet of Things (IIoT)-enabled framework for health monitoring," *Computer Networks*, 2016.

[3] G. Muhammad, "Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system," *Cluster Computing*, vol. 18, no. 2, pp. 795–802, 2015.

[4] M. Shamim Hossain, G. Muhammad, M. F. Alhamid, B. Song, and K. Al-Mutib, "Audio-visual emotion recognition using big data towards 5G," *Mobile Networks and Applications*, 2016.

[5] M. S. Hossain, "Cloud-supported cyber-physical localization framework for patients monitoring," *IEEE Systems Journal*, 2015.

[6] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, M. Alsulaiman, and M. Bukhari, "Formant analysis in dysphonic patients and automatic Arabic digit speech recognition," *BioMedical Engineering Online*, vol. 10, article 41, 2011.

[7] G. Muhammad, M. AlSulaiman, A. Mahmood, and Z. Ali, "Automatic voice disorder classification using vowel formants," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '11)*, pp. 1–6, Barcelona, Spain, July 2011.

[8] M. Bouchayer, G. Cornut, E. Witzig, R. Loire, J. B. Roch, and R. W. Bastian, "Epidermoid cysts, sulci, and mucosal bridges of the true vocal cord: a report of 157 cases," *The Laryngoscope*, vol. 95, no. 9, pp. 1087–1094, 1985.

[9] M. M. Johns, "Update on the etiology, diagnosis, and treatment of vocal fold nodules, polyps, and cysts," *Current Opinion in Otolaryngology and Head and Neck Surgery*, vol. 11, no. 6, pp. 456–461, 2003.

[10] K. Wu and D. G. Childers, "Gender recognition from speech. Part I: coarse analysis," *Journal of the Acoustical Society of America*, vol. 90, no. 4 I, pp. 1828–1840, 1991.

[11] S. M. R. Azghadi, M. R. Bonyadi, and H. Sliahhosseini, "Gender classification based on feedforward backpropagation neural network," in *Artificial Intelligence and Innovations 2007: From Theory to Applications: Proceedings of the 4th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI 2007)*, C. Boukis, L. Pnevmatikakis, and L. Polymenakos, Eds., vol. 247 of *IFIP The International Federation for Information Processing*, pp. 299–304, Springer, Berlin, Germany, 2007.

[12] S. Gaikwad, B. Gawali, and S. C. Mehrotra, "Gender identification using SVM with combination of MFCC," *Advances in Computational Research*, vol. 4, no. 1, pp. 69–73, 2012.

[13] M. Pronobis and M. Magimai-Doss, "Analysis of F0 and cepstral features for robust automatic gender recognition," Tech. Rep. Idiap-RR-30-2009, Idiap, 2009.

[14] Y.-M. Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech," in *Proceedings of the International Conference on Machine Learning and Cybernetics*, pp. 3376–3379, Dalian, China, August 2006.

[15] G. Chen, X. Feng, Y. Shue, and A. Alwan, "On using voice source measures in automatic gender classification of children's speech," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH '10)*, pp. 673–676, Chiba, Japan, 2010.

[16] F. Lingenfelser, J. Wagner, T. Vogt, J. Kim, and E. André, "Age and gender classification from speech using decision level fusion and ensemble based techniques," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH '10)*, pp. 2798–2801, Chiba, Japan, September 2010.

[17] K. Rakesh, S. Dutta, and K. Shama, "Gender recognition using speech processing techniques in labview," *International Journal of Advances in Engineering & Technology*, vol. 1, no. 2, pp. 51–63, 2011.

[18] M. Sigmund, "Gender distinction using short segments of speech signal," *International Journal of Computer Science and Network Security*, vol. 8, no. 10, pp. 159–162, 2008.

[19] D. S. Deiv, Gaurav, and M. Bhattacharya, "Automatic gender identification for hindi speech recognition," *International Journal of Computer Applications*, vol. 31, no. 5, pp. 1–8, 2011.

[20] V. N. Sorokin and I. S. Makarov, "Gender recognition from vocal source," *Acoustical Physics*, vol. 54, no. 4, pp. 571–578, 2008.

[21] M. H. Sedaaghi, "A comparative study of gender and age classification in speech signals," *Iranian Journal of Electrical Electronic & Engineering*, vol. 5, no. 1, pp. 1–12, 2009.

[22] M. Alsulaiman, Z. Ali, and G. Muhammad, "Gender classification with voice intensity," in *Proceedings of the 5th European Modeling Symposium of Mathematical Modeling and Computer Simulation*, pp. 205–209, Madrid, Spain, November 2011.

[23] M. Alsulaiman, Z. Ali, and G. Muhammad, "Voice intensity based gender classification by using simpson's rule with SVM," in *Proceedings of the 19th International Conference on Systems, Signals and Image Processing*, pp. 570–573, Vienna, Austria, April 2012.

[24] I. Mporas, T. Ganchev, E. Kotinas, and N. Fakotakis, "Examining the influence of speech frame size and number of cepstral coefficients on the speech recognition performance," in *Proceedings of the 12th International Conference on Speech and Computer*, pp. 1–6, Moscow, Russia, 2007.

[25] R. Jang, Audio Signal Processing and Recognition: End-Point Detection in Time Domain, March 2016, http://mirlab.org/jang/books/audioSignalProcessing/epdTimeDomain.asp?title=6-2%20EPD%20in%20Time%20Domain%20(%BA%DD%C2I%B0%BB%B4%FA%A1G%AE%C9%B0%EC%AA%BA%A4%E8%AAk).

[26] C. F. Gerald and P. O. Wheatley, *Applied Numerical Analysis*, Pearson, 7th edition, 2003.

[27] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, NY, USA, 9th edition, 1972.

[28] A. Horwitz, "A version of Simpson's rule for multiple integrals," *Journal of Computational and Applied Mathematics*, vol. 134, no. 1-2, pp. 1–11, 2001.

[29] S. Haykin, *Neural Networks a Comprehensive Foundation*, McMaster University, Hamil-ton, Ontario, Canada, 2nd edition, 1998.

[30] S. M. Kamruzzaman, A. N. M. Rezaul Karim, S. Islam, and E. Haque, "Speaker Identification using MFCC-Domain support vector machine," *International Journal of Electrical and Power Engineering*, vol. 1, no. 3, pp. 274–278, 2007.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM," Tech. Rep., NIST, 1993.

[32] M. Alsulaiman, G. Muhammad, M. A. Bencherif, A. Mahmood, Z. Ali, and M. Aljabri, "Building a rich arabic speech database," in *Proceedings of the 5th Asia Modeling Symposium (AMS '11)*, pp. 100–105, Kuala Lumpur, May 2011.

[33] M. A. Anusuya and S. K. Katti, "Front end analysis of speech recognition: a review," *International Journal of Speech Technology*, vol. 14, no. 2, pp. 99–145, 2011.

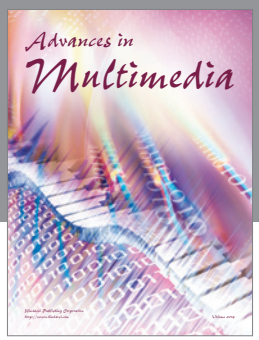
[34] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

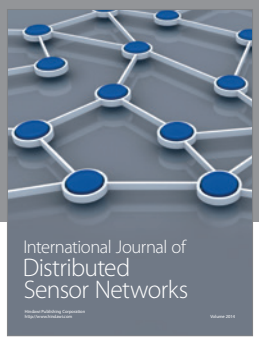
[35] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Upper Saddle River, NJ, USA, 1988.

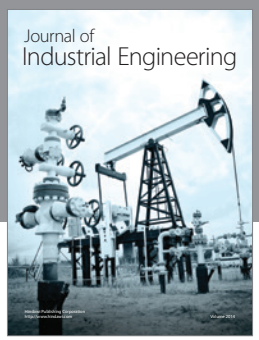
[36] J. I. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, and P. Gómez-Vilda, "Automatic detection of voice impairments from text-dependent running speech," *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 176–182, 2009.

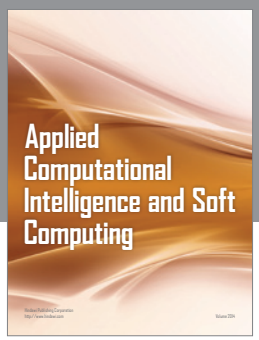
[37] M. S. Hahn, B. A. Teply, M. M. Stevens, S. M. Zeitels, and R. Langer, "Collagen composite hydrogels for vocal fold lamina propria restoration," *Biomaterials*, vol. 27, no. 7, pp. 1104–1109, 2006.

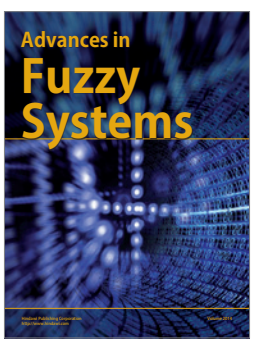
[38] I. R. Titze, *Principles of Voice Production*, Prentice Hall, 1st edition, 1994.

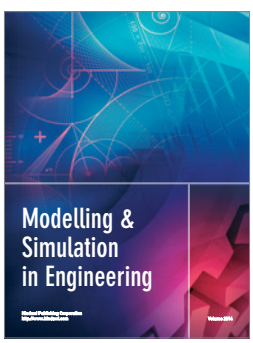


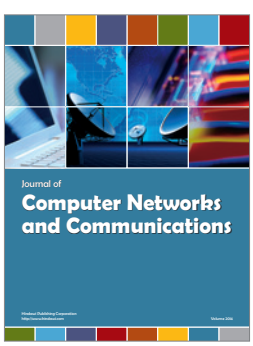












Hindawi
Submit your manuscripts at
http://www.hindawi.com